**Feature Engineering with Analytical Data**

**Industry Mentors (Dow Chemical)**

Ben Reiner, Christian Heil, Lyndsay Leal,

Maria C. Folgueras, Mingkang Sun, Dan Miller, Steve Arturo, Alix Schmidt

**Authors (University of Delaware)**

Suhas Cristy Mathey, M.S., Computer Engineering

Juliana Nam, 2nd year Ph.D. Candidate, Chemical & Biomolecular Engineering

**CHEG/CISC/ECEG/MSEG 848**

Professor Arthi Jayaraman

Vinsion Liao, Graduate Student, Chemical & Biomolecular Engineering

Spring 2024

**Abstract**

The goal of this project is to create a computational workflow that takes in input analytical characterization data, for various materials of interest to Dow Chemical, and generates a quantitative polymer structure-performance relationship. The raw data provided by Dow Chemical included three forms of analytical characterization data obtained from NMR, FTIR, and GPC experiments for 12 material samples. We pre-processed this data to reduce the amount of inherent noise from the experimental measurement and instrumentation. We then take the pre-processed data as input to principal component analysis to select features and reduce the dimensionality of the dataset for it to be used as input for models evaluating the correlation of this data to material performance. The key outcomes from our work in this project have been various rigorous pre-processing methods, feature analysis process, and two linear regression models.

**Introduction**

Dow is a materials science company that offers a wide range of products and services, including commercial personal care products. However, with the increase in corporate sustainability, Dow also seeks to create more sustainable and inclusive personal care solutions. This change requires more greener chemistries in their commercial products by replacing synthetic polymers with bio derived polymers, which will ultimately increase the biodegradability of their products to comply with ISO 16128. However, to change formulations with greener chemistries would require numerous creations of new formulations and testing operations of said formulations. This can be time consuming and involve many variables, especially revolving around the analysis of compositions. One way of analyzing different compositions and samples is with spectroscopy measurements, such as nuclear magnetic resonance (NMR) spectroscopy, fourier transform infrared (FTIR) spectroscopy, and gel permeation chromatography (GPC). Traditionally, specific NMR chemical shifts and peak intensities are observed to identify specific structural properties in a molecule;1 certain FTIR absorbance peaks and their respective peak heights are used to identify molecular compounds;2 and GPC peak heights and widths over a period of retention time are used to determine the relative molecular weight of the sample and the molecular weight distribution of compounds that make up the sample.3 However, the traditional method for analyzing spectroscopy data varies and is subjective, where only specific peak values or positions are taken into consideration. Therefore, Dow requires a rigorously reproducible program that will intake all of the analytical data, including all peaks and peak positions, and output its results without bias.

**Project Goal**

Dow provided 12 samples of three different types of spectroscopy data (NMR, FTIR, and GPC). The goal of this project is to create a data driven model to correlate polymer features, that are extracted from the measurements, to a performance metric that has also been provided, where each sample has been assigned a performance value, as seen in Appendix A. The steps to accomplish this task is an iterative process where the data will be pre-processed and transformed for feature extraction and put through a forward model to correlate features to performance. Therefore, the objective of this project is to leverage analytical characterization data to create an analytically informed hybrid modeling strategy, with a focus on creating rigorous methods for featurization of analytical spectra.

**Computational Approach**

The project has been broken down into four main tasks: pre-processing, feature selection, selection and application of models, and quantification and correlation of specific features to performance.4

**Pre-Processing**

For data driven models, pre-processing is needed to make raw data more interpretable for the eventual model application.5 In this case, there is uncertainty in the data due to inherent noise from the experiment instrumentation and overall experimental process. Additionally, there are only 12 samples, but each individual dataset is made up of thousands of points, resulting in a high dimensional raw dataset. Therefore, pre-processing is required not only to eliminate noise but also reduce the dimensionality of the dataset before introducing it to the model.

**Feature Selection**

To further reduce the dimensionality, feature selection was applied to determine the most important features to translate to the eventual model.6 As all of the three types of spectroscopy data are plots that contain curves and respective peaks, peak intensity was used as the criteria for feature selection. As for the method for feature selection, Principal Component Analysis (PCA) was chosen. PCA is a common linear dimensionality reduction technique that assists in simplifying the complexity of high-dimensional data.7 It reduces the feature space by creating new variables (principal components) that are linear combinations of the original features: peaks.

**Model Selection and Application**

To select an appropriate model for analyzing PCA reduced data, machine learning and traditional statistical models, such as linear regression, were considered. However, a key characteristic of the dataset is the low sample size of the dataset, for there are only 12 samples. Complex machine learning models traditionally require large datasets to produce results with high accuracy. Therefore, a simple linear model was chosen to utilize the low sample size of the dataset while reducing the risk of overfitting. Linear regression models provide clear and direct relationships between input features (principal component 1 and principal component 2) and the output (performance metric).

**Quantification and Correlation of Specific Features to Performance**

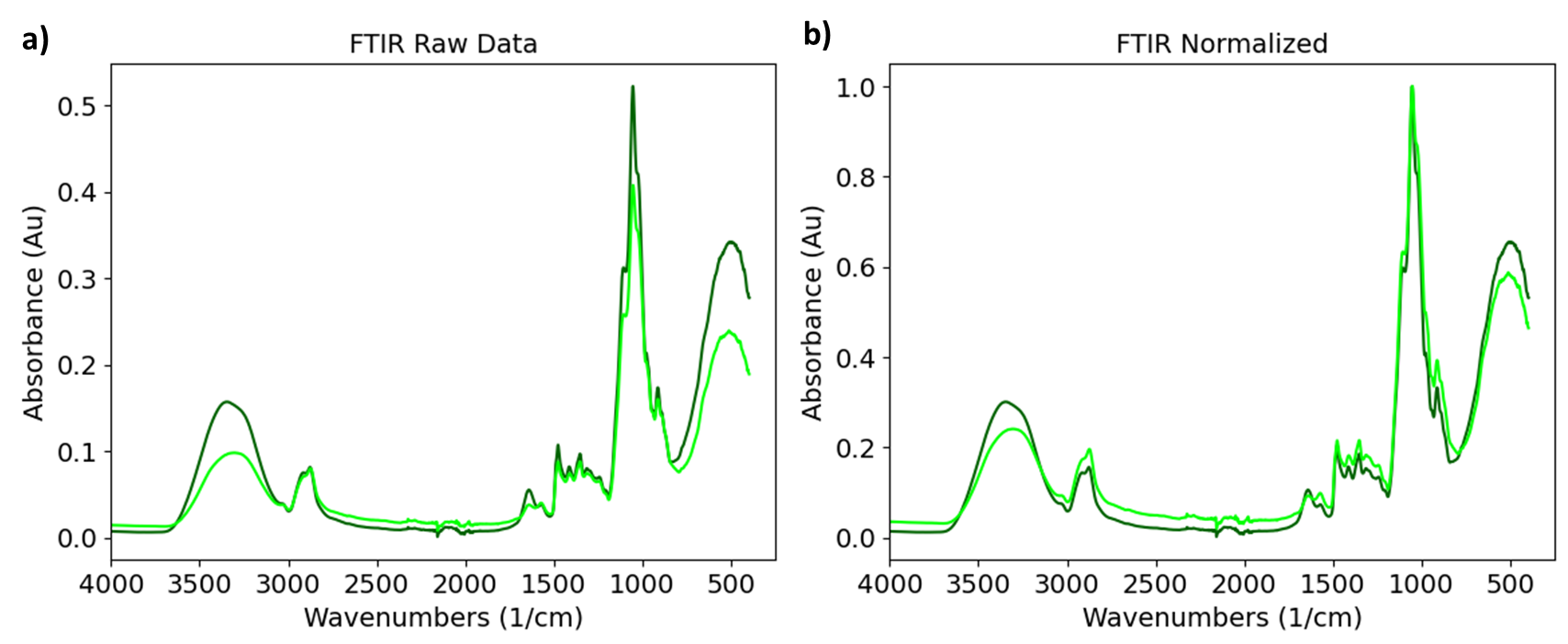
The final task was unable to be completed due to lack of time; however, the goal of this task is to determine specific features that greatly affect performance and quantify their direct relationship to performance.

**Results**

**Pre-Processing**

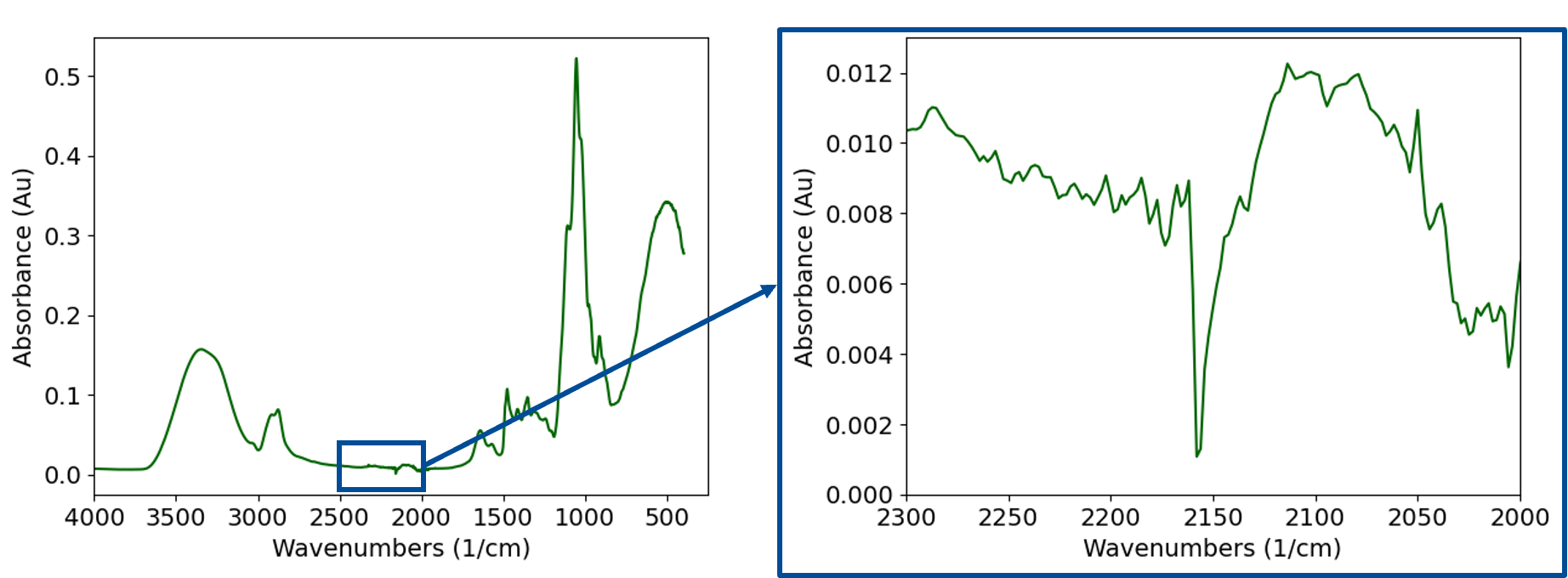
There are many methods of pre-processing, such as normalizing, binning, first and second derivative, etc.;5 however, this report will detail the pre-processing methods for which rigorous methods were created. The two main methods of pre-processing, in the scope of this project, are normalizing and binning.

Normalization rescales the dataset with itself to make the datasets more comparable to each other.8 For example, Figure 1a depicts two samples that appear to be similar where the darker curve is greater in magnitude compared to the lighter curve. However, when it has been normalized in Figure 1b, the two plots are put on an equal scale and can be compared more holistically. This can be helpful not only to compare across samples but also between different spectroscopy measurements (ex: FTIR vs GPC).



**Figure 1.** Example of Normalization. a) non-normalized FTIR data. b) normalized FTIR data.

The second pre-processing method that was applied to the provided spectroscopy data is binning: the process of dividing the provided dataset into intervals known as bins to lower the dimensionality of the dataset.10 At small length scales, there is a lot of fluctuation in between points, and this is only emphasized by the large size of each dataset. This is due to the noise from instrumentation and generally from the experimental process. This is depicted by Figure 2, where an enlarged view of an FTIR sample shows a lot of noise and fluctuation in the noise region from 2000 to 2300 wavenumbers.



**Figure 2.** Visualization of Noise in Example FTIR Plot.

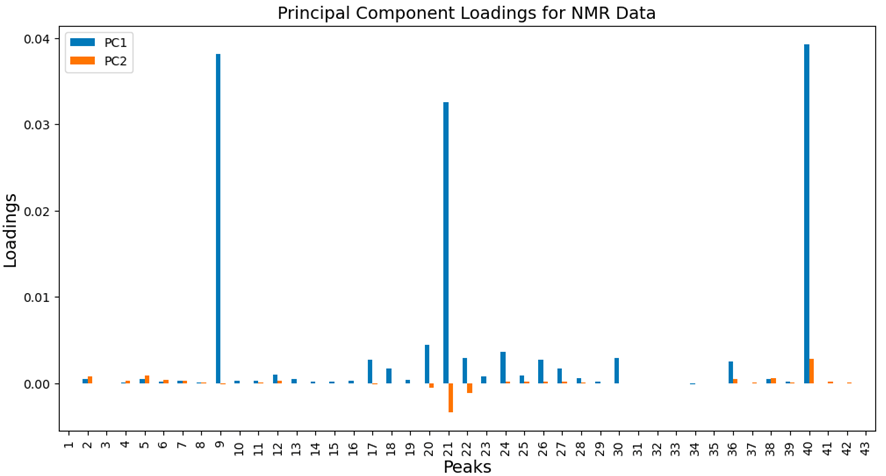
Furthermore, there are multiple types of binning methods, such as naive binning and dynamic binning. Naive binning involves having a uniform range of data points per bin; the dataset gets evenly divided and the average of the curve for each bin is compiled. Dynamic binning involves compiling the average values of a peak per bin. To accomplish this, an existing peak locating algorithm is utilized to determine peak location and width to then separate all peaks into their own bins. The value of binning is to eliminate noise, which we can see an example of in Figure 2, and reduce dimensionality due to the process of averaging a range of points in a bin to then represent them as a single value. For example, dynamic binning on GPC data can turn about 8,000 data points into only a couple of peaks.

Normalization, naive binning, and dynamic binning are all data driven pre-processing methods that can be applied depending on sample chemistry and inputted spectroscopy data. Although the goal is to turn away from traditional analytical methods, it would be unwise to not incorporate expert knowledge. For example, Dow noted that the large peak in all NMR datasets, represented by the large blue peak in Figure 3, is a water or solvent peak that is unrelated to the samples and is only apparent due to the experimental process of obtaining NMR data. Therefore, expert knowledge was incorporated to eliminate the water peak, leading to the emphasis of related data, represented by the orange spectrum in Figure 3. This shows that while data driven methods are objective and align with the goal to create holistic analytical methods, expert driven methods are still needed to create well performing and well rounded models.

**Feature Selection**

After preprocessing the spectroscopy data, feature selection is required to further reduce the dimensionality of the data and determine the most important features to send to the model. Principal component analysis is able to turn peaks from dynamic binning or average values from naive binning to only two principal components.

One set of results from PCA is a scatter plot that displays how all 12 samples are distributed based on PCA analysis; an example shown in Figure 4 is for all 12 NMR samples. Conclusions can also be drawn from the scatterplot. For example, the positioning for NMR is intriguing due to the isolated sample 2 in the top right corner. It just so happens that sample 2 also exhibits the highest mean performance, which could suggest a potential connection between sample 2’s NMR spectral features and enhanced performance. Understanding these patterns can help to predict how specific features might influence the overall performance, guiding future modifications or selections particularly in the last task, quantification and correlation of specific features to performance.



**Figure 5.** Principal Component Loadings for NMR Data.

Another feature of the combination of PCA and binning is to determine the bins or, in the case of dynamic binning, the peaks that contribute to each principal component. Figure 5 depicts the principal component loadings for all 12 NMR samples and the correlation between peaks and principal components. These loadings indicate the magnitude of contribution of each peak to each principal component. A higher absolute value of the loadings means the corresponding peak has a significant influence over a principal component. In the example shown in Figure 5, it is shown that particularly higher loadings were observed for Peaks 9, 21, and 40, meaning these peaks’ variance were captured in principal component 1. 

The reason only two principal components were chosen for feature selection is shown in Figure 6, a scree plot. A scree plot is a visual tool that displays the explained variance against the number of components to determine the required number of components. As seen in Figure 6, the variance significantly drops at the 2nd component level, indicating an optimal cutoff point for analysis. Therefore, selecting two principal components maintains a balance between simplifying the model and capturing the essential variability in the data, allowing for more targeted and interpretable results.

**Model Selection and Application**

Traditional statistical models reduce the risk of overfitting with the lower sample size dataset that was provided. Therefore, linear regression and polynomial regression are applied to test the translation of PCA to a regression model.

Table 1 displays the performance of the linear regression for dynamically binned data, where the values are R2 values and explains how well the variability in the performance data is accounted for by the model. A higher R2 value indicates a better fit of the model to the data. As seen in Table 1, the individual datasets resulted in a low performing model; however, applying linear regression for combined principal components of NMR, FTIR, and GPC results in a high performing model, demonstrating the value of a comprehensive model that integrates multiple data sources. Linear regression resulting R2 values for both dynamically and naively binned are available in Appendix F. Table 1 also displays the performance of the polynomial regression, which shows improved model results for individual data sets. By introducing polynomial terms, non-linear relationships between principal components and performance is allowed. This explains the improvement, for polynomial regression extends linear models byu adding terms that allow for a curved relationship between the independent and dependent variables. However, polynomial regression is unable to be applied on the combined dataset for the number of parameters in the regression equation will be greater than the number of samples, resulting in an infinite number of solutions. Furthermore, these results can also be explained by parity plots, which visually depict the performance by plotting predicted and actual performance. These plots can be found in Appendix F and G for naive and dynamic binning, respectively.

**Table 1.** Regression results for naive and dynamic binning for both linear and polynomial regression

|  | Linear Regression (deg = 1)  R2 values | | Polynomial Regression (deg = 2)  R2 values | |
| --- | --- | --- | --- | --- |
|  | Naive | Dynamic | Naive | Dynamic |
| NMR | 0.225 | 0.272 | 0.559 | 0.405 |
| FTIR | 0.047 | 0.055 | 0.658 | 0.616 |
| GPC | 0.374 | 0.373 | 0.837 | 0.840 |
| Combined (NMR, FTIR, GPC) | 0.467 | 0.804 | - | - |

**Conclusion**

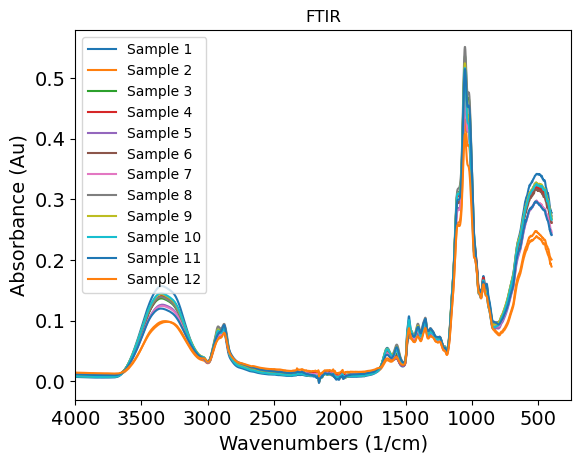
The work completed over the past few months include multiple user friendly pre-processing methods for spectroscopy data: normalization, naive binning, dynamic binning, finding peaks, and implementation of expert knowledge; a feature selection method: principal component analysis (PCA); and a couple regression models that can be applied to the previously extracted features to correlate to performance. There is still room for improvement in regard to the performance of the model, but this can be done with implementation of different combinations of datasets, more pre-processing methods, or different types of models, such as polynomial, logistic, or lasso regression. However, once the model is able to perform well with the provided features, the determination of the features that directly impact performance will be useful for Dow’s formulation process.

**References**

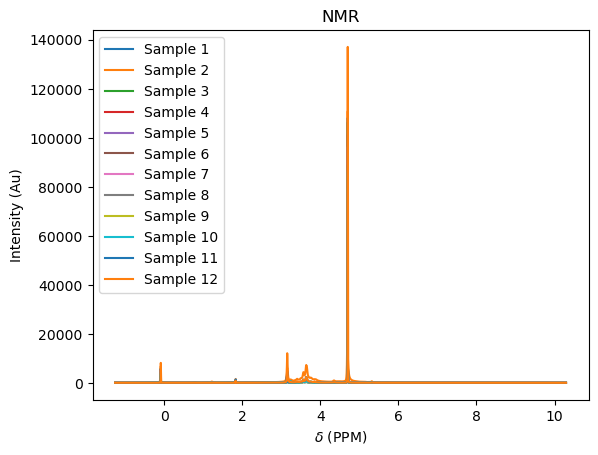
1. Reusch, W. *Nuclear Magnetic Resonance Spectroscopy*. Chemistry LibreTexts. <https://chem.libretexts.org/Bookshelves/Organic_Chemistry/Supplemental_Modules_(Organic_Chemistry)/Spectroscopy/Nuclear_Magnetic_Resonance_Spectroscopy>.
2. *William.R.Stockwell*. Physical Chemistry. <https://chem.libretexts.org/Courses/Howard_University/Howard%3A_Physical_Chemistry_Laboratory/14%3A_Fourier_Transform_Infrared_Spectroscopy_(FTIR)>.
3. Moore, J. C. Gel Permeation Chromatography. I. A New Method for Molecular Weight Distribution of High Polymers. *Journal of Polymer Science Part A: General Papers* 1964, *2* (2), 835–843. https://doi.org/10.1002/pol.1964.100020220.
4. *What is Chemometrics and Why Study it?* Chemistry LibreTexts. <https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Chemometrics_Using_R_(Harvey)/00%3A_Front_Matter/What_is_Chemometrics_and_Why_Study_it%3F>.
5. Al-jabery, K. K.; Obafemi-Ajayi, T.; Olbricht, G. R.; Wunsch II, D. C. *2 - Data preprocessing*. ScienceDirect. <https://www.sciencedirect.com/science/article/abs/pii/B9780128144824000024>.
6. *Chemometrics-based Spectroscopy for Pharmaceutical and Biomedical Analysis | Frontiers Research Topic*. www.frontiersin.org. https://www.frontiersin.org/research-topics/6301/chemometrics-based-spectroscopy-for-pharmaceutical-and-biomedical-analysis/articles (accessed 2024-05-13).
7. Beattie, J. R.; Esmonde-White, F. W. L. Exploration of Principal Component Analysis: Deriving Principal Component Analysis Visually Using Spectra. *Applied Spectroscopy* 2021, 000370282098784. <https://doi.org/10.1177/0003702820987847>.
8. Li, B.; Wu, F.; Lim, S.-N.; Belongie, S.; Weinberger, K. Q. *On Feature Normalization and Data Augmentation*. openaccess.thecvf.com. https://openaccess.thecvf.com/content/CVPR2021/html/Li\_On\_Feature\_Normalization\_and\_Data\_Augmentation\_CVPR\_2021\_paper.html?ref=https://githubhelp.com.

**‌**

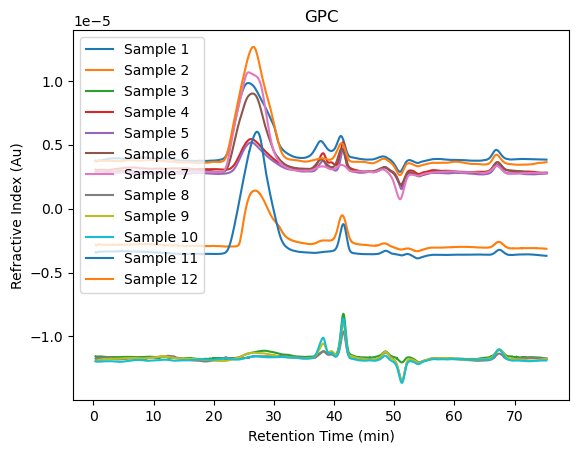
**Appendix A: Provided Data (Raw FTIR, NMR, and GPC data + Performance Metric)**



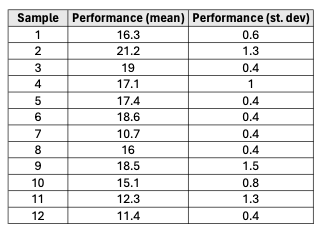
**Figure A1.** FTIR Raw Data for All 12 Samples.



**Figure A2.** NMR Raw Data for All 12 Samples.

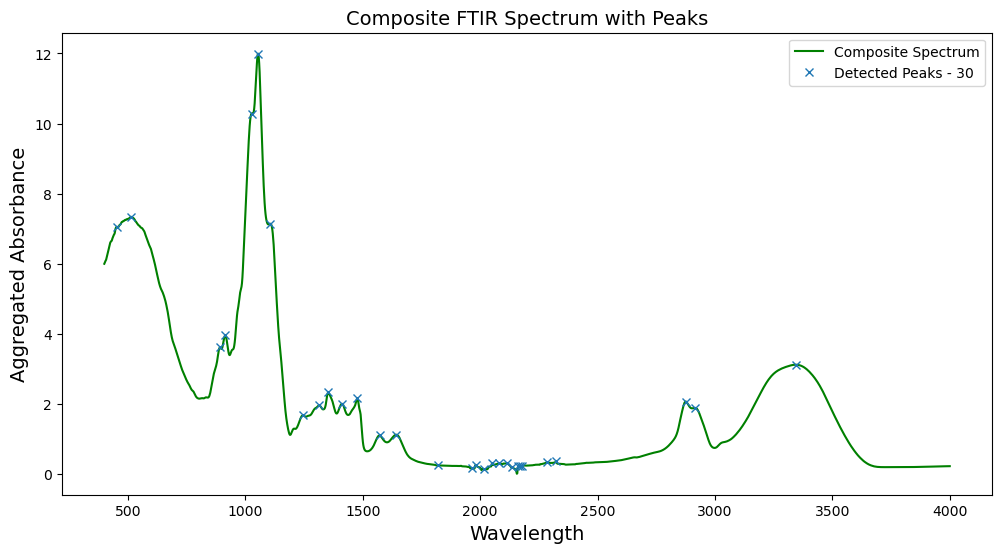


**Figure A3.** GPC Raw Data for All 12 Samples.

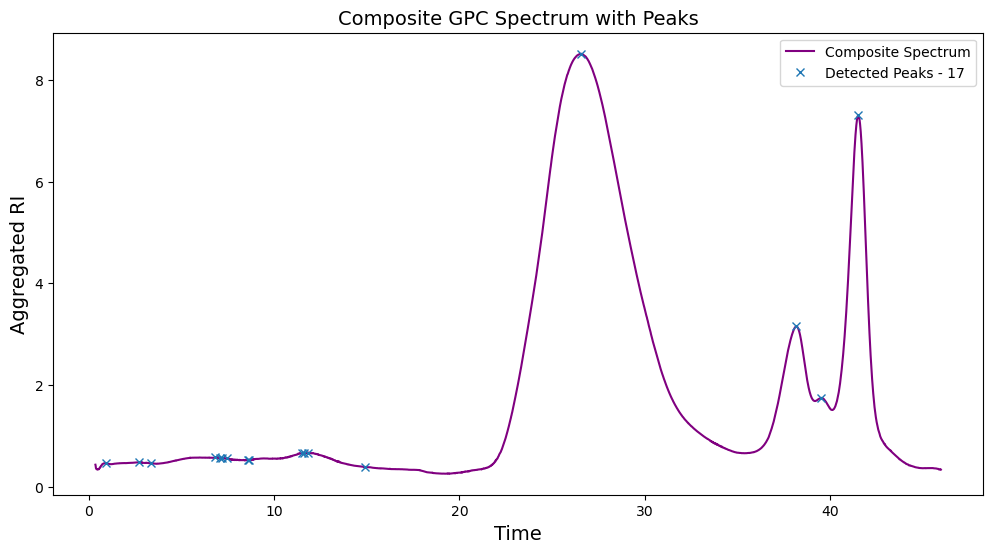


**Figure A4.** Performance Metric for All 12 Samples.

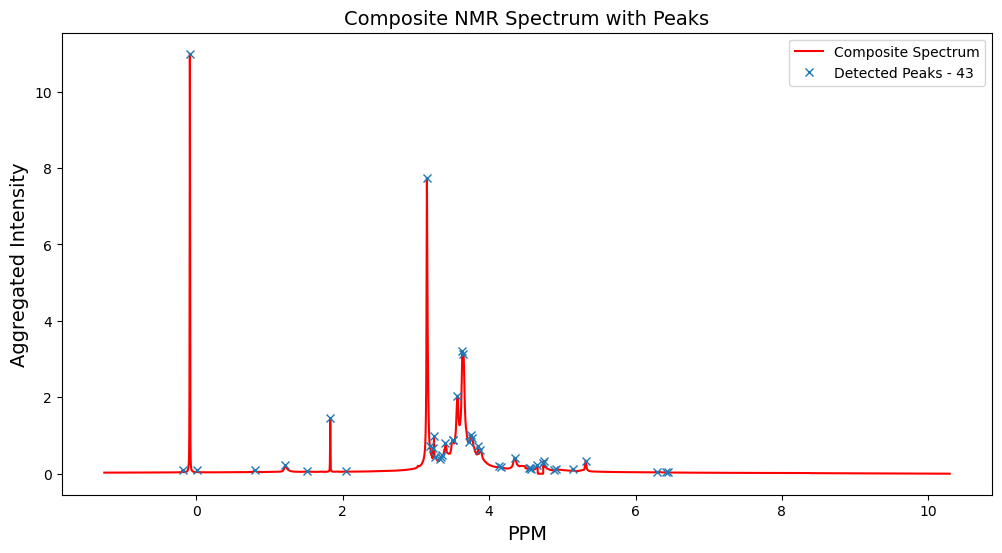
**Appendix B: Peak Detection for Dynamic Binning**

****

**Figure C1.** Composite Spectrum with Peak Detection for FTIR.

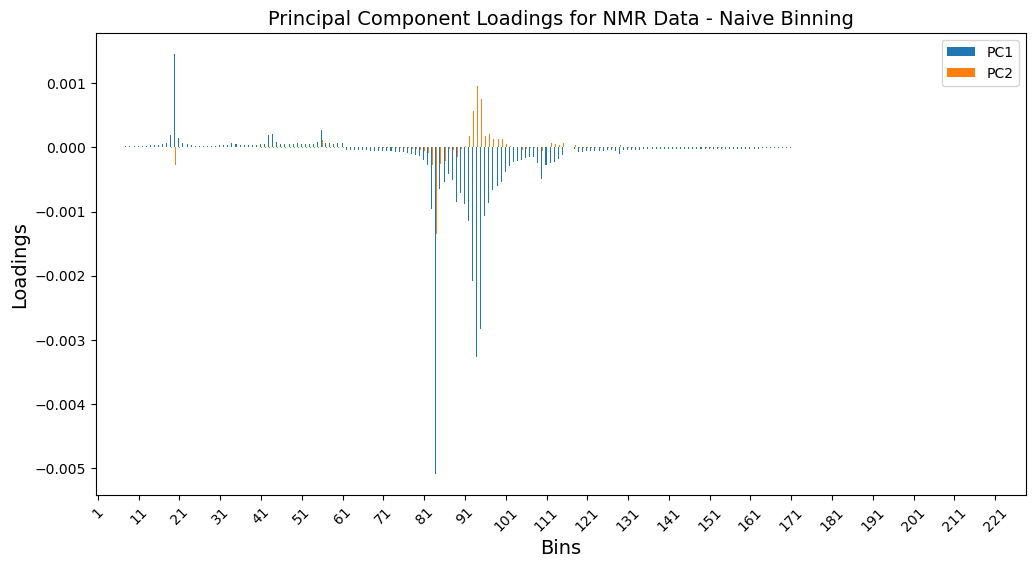
****

**Figure C2.** Composite Spectrum with Peak Detection for GPC.

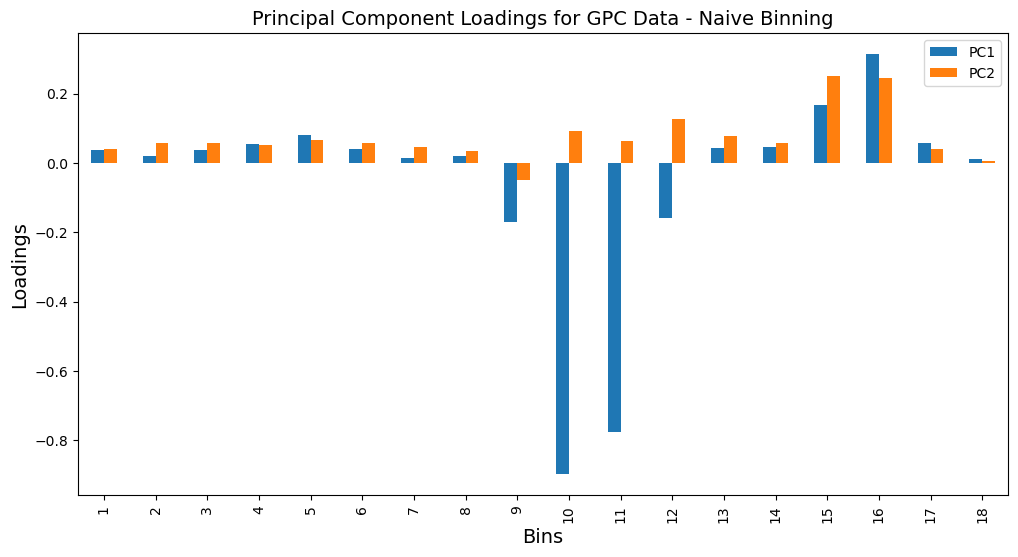
****

**Figure C1.** Composite Spectrum with Peak Detection for NMR.

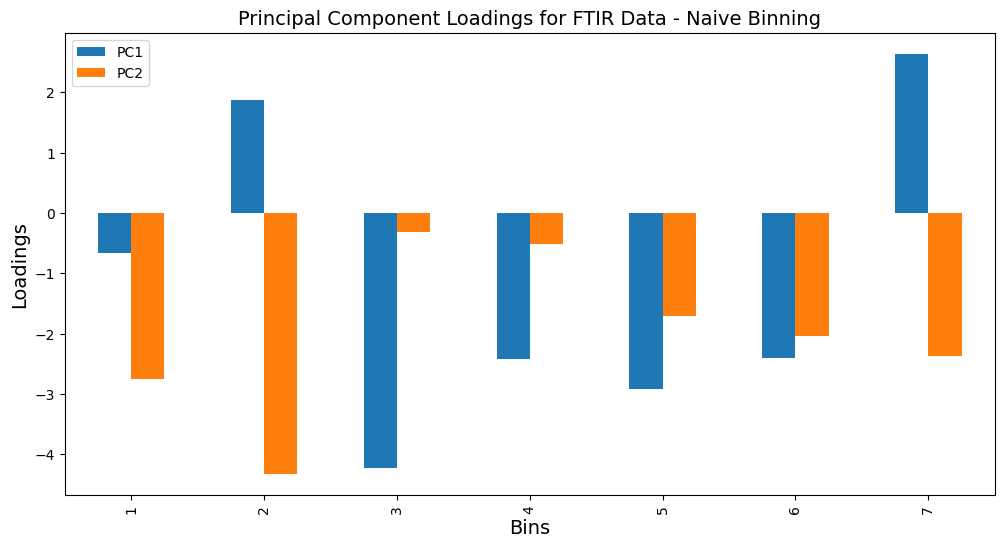
**Appendix D: Principal Component Loadings for NMR, FTIR, and GPC**

****

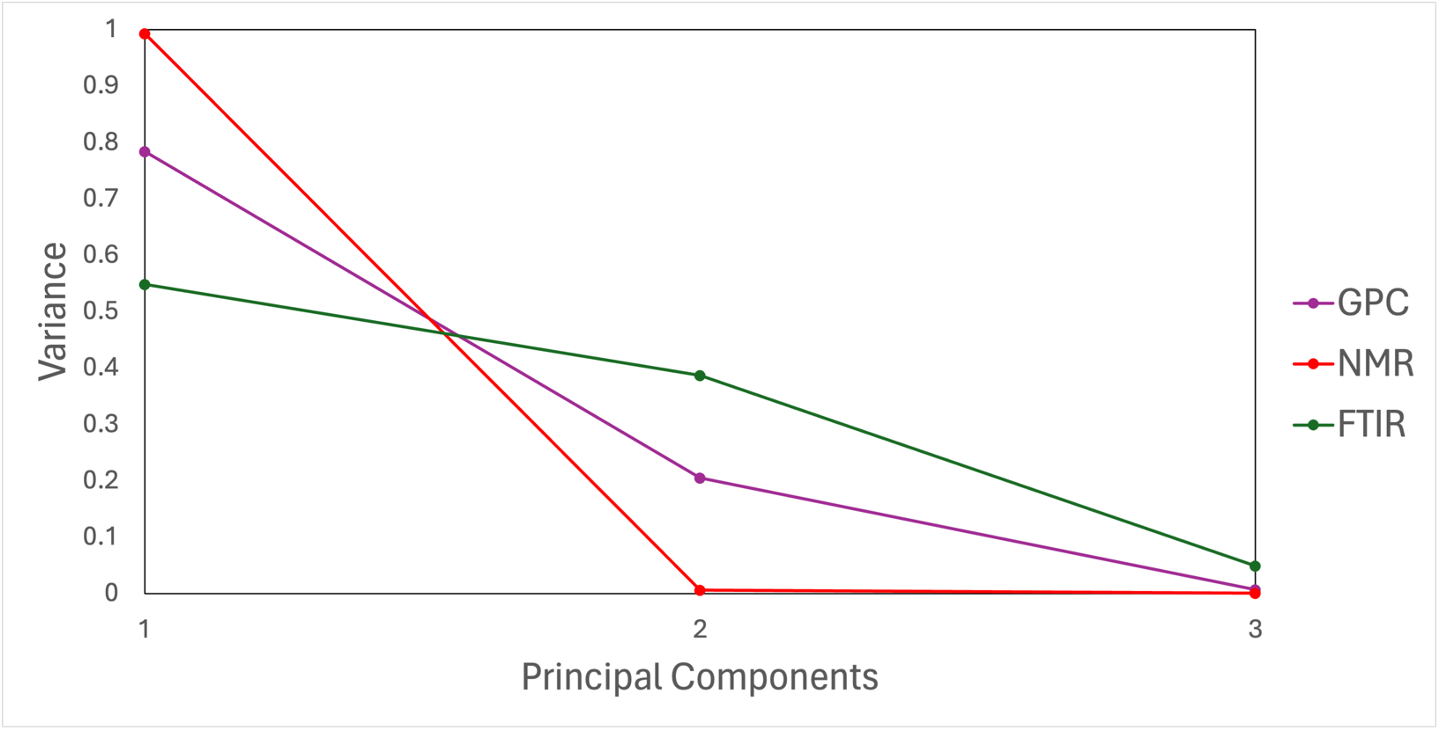
**Figure D1.** Principal Component Loadings Plot for Naively Binned NMR Data.

****

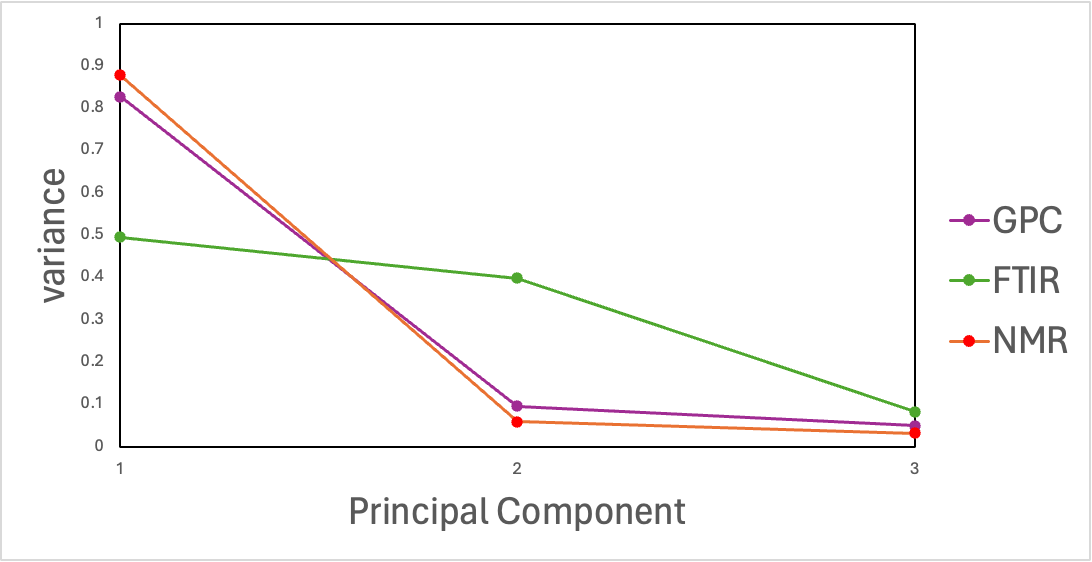
**Figure D2.** Principal Component Loadings Plot for Naively Binned GPC Data.

**Figure D3.** Principal Component Loadings Plot for Naively Binned FTIR Data.

**Appendix E: Scree Plots for Dynamically and Naively Binned Data**

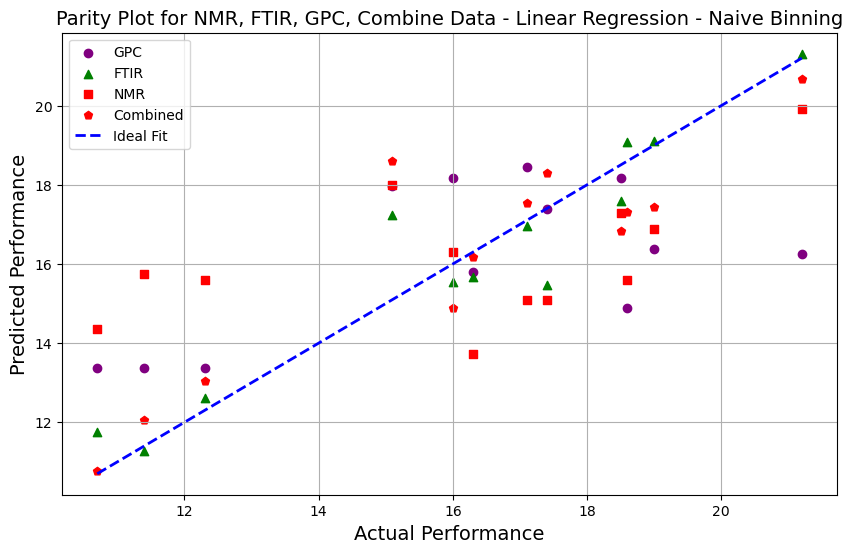
****

**Figure E1.** Scree Plot for Dynamically Binned Data

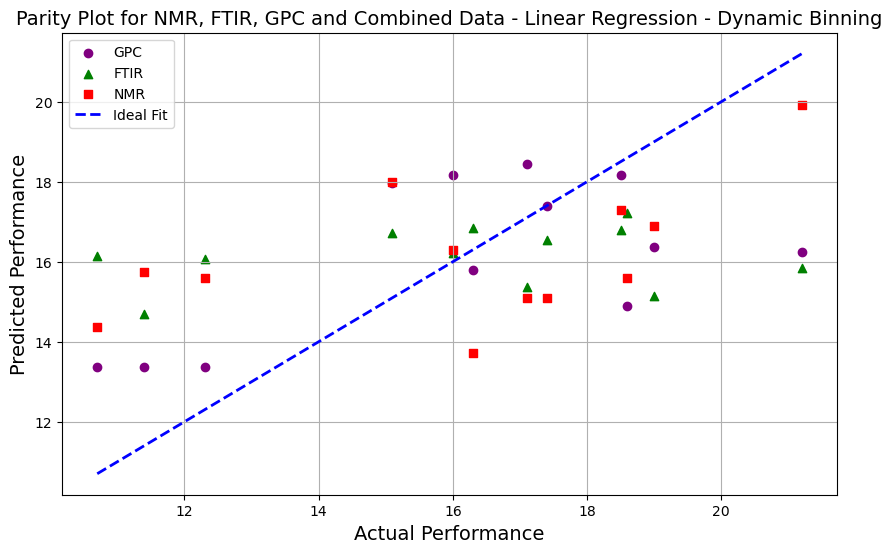
****

**Figure E2.** Scree Plot for Naively Binned Data

**Appendix F: Linear Regression Results**

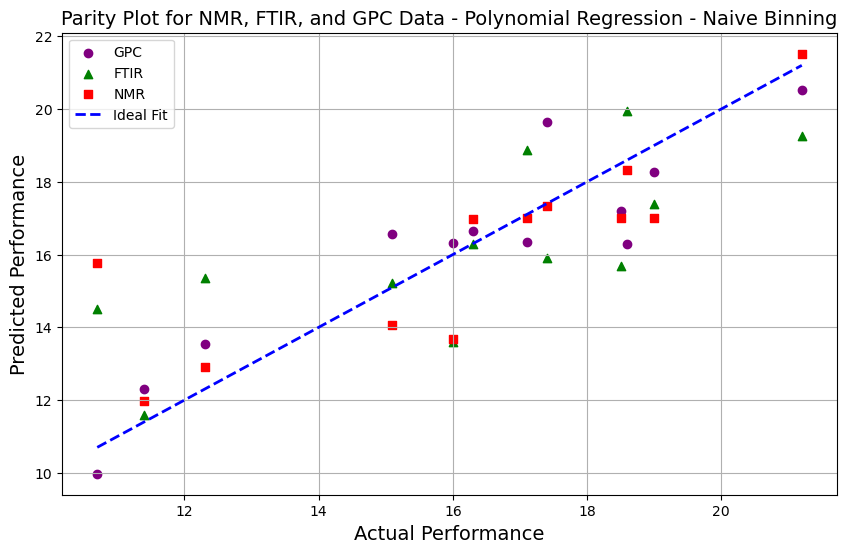
****

**Figure F1.** Parity Plot for Linear Regression on Naively Binned Data.

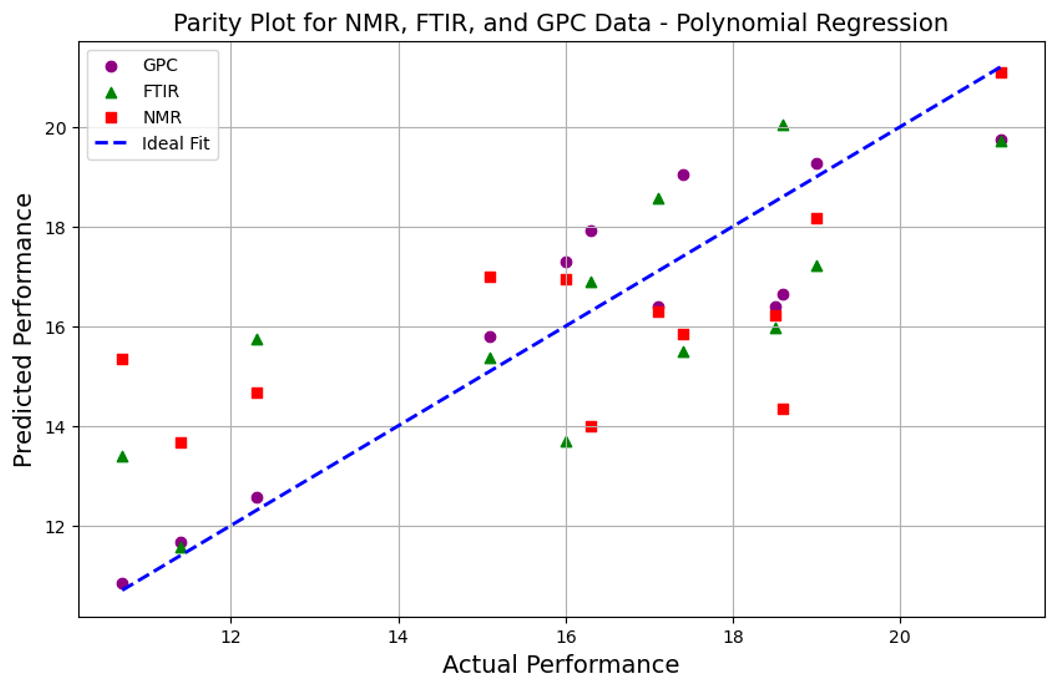


**Figure F2.** Parity Plot for Linear Regression on Dynamically Binned Data.

**Appendix G: Polynomial Regression results**

****

**Figure G1.** Parity Plot for Polynomial Regression on Naively Binned Data.



**Figure G2.** Parity Plot for Polynomial Regression on Dynamically Binned Data.